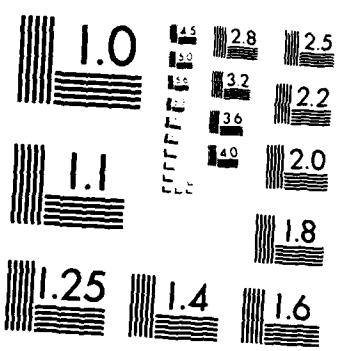
RD-A129 725 TIME SERIES MODEL IDENTIFICATION BY ESTIMATING INFORMATION(U) TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS E PARZEN NOV 82 TR-N-35 1/1

UNCLASSIFIED N00014-82-MP-2001 F/G 12/1 NL



TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS 77843-3143

INSTITUTE OF STATISTICS
Phone 713 - 845 3141

# TIME SERIES MODEL IDENTIFICATION BY ESTIMATING INFORMATION



by Emanuel Parzen

Institute of Statistics

Texas A&M University

Technical Report No. N-35

November 1982

Texas A&M Research Foundation
Project No. 4226T

"Multiple Time Series Modeling and
Time Series Theoretic Statistical Methods"

Sponsored by the Office of Naval Research

Professor Emanuel Parzen, Principal Investigator



Approved for public release; distribution unlimited.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>N-35 | 2. GOVT ACCESSION NO.<br>AD-A129 725 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>Time Series Model Identification by Estimating Information | | 5. TYPE OF REPORT & PERIOD COVERED<br>Technical<br>6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>Emanuel Parzen | | 8. CONTRACT OR GRANT NUMBER(s)<br>ONR N00014-82-MP-2001 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Texas A&M University<br>Institute of Statistics<br>College Station, TX 77843 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS |
| 11. CONTROLLING OFFICE NAME AND ADDRESS | | 12. REPORT DATE<br>Nov. 1982<br>13. NUMBER OF PAGES |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>Unclassified<br>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Time series analysis, identification, information, entropy, feedback, ARMA scheme, multiple time series analysis.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number) Statisticians, economists, and system engineers are becoming aware that to identify models for time series and dynamic systems, information theoretic ideas can play a valuable (and unifying) role. This paper discusses how models for a univariate or multivariate time series Y(t) can be formulated as hypotheses about the information divergence between alternative models, formulated as sets of variables that are sufficient to forecase Y(t). These information numbers play a central role in studies of causality and feedback.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

# TIME SERIES MODEL IDENTIFICATION BY ESTIMATING INFORMATION

by

Emanuel Parzen

Institute of Statistics
Texas A&M University



Dedicated to Professor Theodore W. Anderson
in Celebration of his 65th Birthday

## ABSTRACT

Statisticians, economists, and system engineers are becoming aware that to identify models for time series and dynamic systems, information theoretic ideas can play a valuable (and unifying) role. This paper discusses how models for a univariate or multivariate time series $Y(t)$ can be formulated as hypotheses about the information divergence between alternative models for the conditional probability density of $Y(t)$ given various bases involving past, current, and future values of $Y(\cdot)$ and related time series $X(\cdot)$. To determine sets of variables that are sufficient to forecast $Y(t)$, and thus to determine a model for $Y(t)$, an approach is presented which estimates and compares various information increments. These information numbers play a central role in studies of causality and feedback. Approximating autoregressive schemes are used to form estimators of the many information numbers that one might compare to identify models for a time series.

Research supported by Office of Naval Research under contract no. N00014-82-MP-20001

## 0. Introduction

In applications of statistical theory, it is important to distinguish between the problem of parameter estimation (which belongs to confirmatory statistical theory) and the problem of model identification (which belongs to exploratory statistical theory). The modeling problem arises in conventional (static) statistics whenever the researcher's goal is to screen variables (that is, to determine which variables (for which measurements exist) are most associated with specified variables which we seek to explain, forecast, or control). Researchers are becoming aware [see IFAC (1982)] that to identify models for time series and dynamic systems, information theoretic ideas can play a valuable (and unifying) role [see Akaike (1977)]. The thrust has been clearly articulated, but how to carry it out has not been clear. That entropy ideas have a role in spectral estimation is being widely stated; however, in my view the nature of the role is not well understood by most users of spectral estimation techniques. This paper does not discuss entropy-based spectral estimation [see Parzen (1982)]; it is concerned with identifying time domain models for univariate and multivariate time series by estimating suitable information measures. Most of the calculations proposed are in the time domain. But spectral density concepts and calculations are also used.

Section 1 states the definition of various information measures for probability densities and for random variables. The conjectured ease of calculating significance levels for tests of hypotheses by estimating information increments is illustrated for the problem of testing independence of normal random variables using sample correlation coefficients.

The formulation of tests for white noise and ARMA models in terms of information measures is discussed in sections 2 and 3. Multiple time series identification is discussed in section 4, and illustrated by an example in section 5.

Analysis of empirical time series using the information measures discussed in this paper has been implemented in our computer subroutine library TIMESBOARD of time series analysis programs which is the creation of Professor H. J. Newton. I would like to express my appreciation to Dr. Newton for his close collaboration in this research program. The work of Parzen and Newton (1980) provides a foundation for section 4 of this paper.

## 1. Role of Information Measures in Model Identification

The concept of information theory most familiar to statisticians is the entropy, denoted H(f), of a continuous distribution with probability density f(x), $-\infty<x<\infty$, defined by [log is taken with base e]

$$H(f) = \int_{-\infty}^{\infty} \{- \log f(x)\} f(x)\, dx.$$

A more general concept is information divergence I(f;g) of a density g(x), usually representing a model, from a density f(x), usually representing the true density. We define

$$I(f;g) = \int_{-\infty}^{\infty} \{-\log \frac{g(x)}{f(x)}\} f(x)\, dx.$$

To express information divergence in terms of entropy, define the cross-entropy H(f;g) of $f(\cdot)$ and $g(\cdot)$ by

$$H(f;g) = \int_{-\infty}^{\infty} \{-\log g(x)\} f(x)\, dx.$$

Information-divergence has the important decomposition

$$(*) \qquad 0 \leq I(f;g) = H(f;g) - H(f).$$

There is an important relation between entropy and measures of deviation (scale parameter) denoted $\sigma$. A location-scale parameter model for a density f(x) is

$$f(x) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right)$$

where $f_0(\cdot)$ is a known density, and $\mu$ and $\sigma$ are parameters to be estimated. One may verify that

$$H(f) = \log \sigma + H(f_0)$$

For a normal distribution, the standard density $f_0(x)$ is usually defined by

$$f_0(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp - \frac{1}{2} x^2;$$

then $H(f) = \log \sigma + \frac{1}{2}\{1 + \log 2\pi\}$. A new standardization of the normal distribution proposed by Stigler (1982) is the density

$$f_0(x) = e^{-\pi x^2} \quad .$$

Then $H(f_0) = 0.5$, and $H(f) = \log \sigma + 0.5$.

One of the aims of this paper is to point out that many familiar statistics for testing hypotheses about the models fitting data can be formulated as entropy-difference statistics. Thus an F-test forms

$$F = \hat{\sigma}_1^2 \div \hat{\sigma}_2^2$$

where $\hat{\sigma}_j^2$ is an estimator of a variance $\sigma^2$ of a normal distribution.

Instead of F, consider Fisher's original proposal to form

$$Z = \frac{1}{2} \log F = \log \hat{\sigma}_1 - \log \hat{\sigma}_2$$

We can write $Z = \hat{H}_1 - \hat{H}_2$, where $\hat{H}_j$ is an estimator of entropy based on $\hat{\sigma}_j^2$. In words, Z is a <u>difference</u> of two different estimators of entropy. Our aim in this paper is to systematically develop statistics for testing model identification hypotheses which can be interpreted as entropy-difference statistics. The entropy-difference statistics that arise in time series can be further interpreted as measuring information. We outline various facts which justify a conjecture that information-based test statistics have similar distributions.

We next define information measures for random variables and time series. For a continuous random variable Y with probability density $f_Y(y)$, the <u>entropy</u> of Y is defined by

$$H(Y) = H(f_Y)$$

For a continuous random variable Y and continuous random vector X the <u>conditional entropy</u> of Y given X is defined

$$H(Y|X) = H(f_{Y|X}) = E_X\, H(f_{Y|X})$$

Explicitly, when X is a random variable,

$$E_X\, H\,(f_{Y|X}) = \int_{-\infty}^{\infty} H(f_{Y|X=x})\, f_X\,(x)\, dx$$

where

$$H(f_{Y|X=x}) = \int_{-\infty}^{\infty} \{-\log f_{Y|X=x}(y)\} f_{Y|X=x}(y)\, dy$$

The information $I(Y|X)$ about a continuous random variable Y in a continuous random variable X is defined by

$$\begin{aligned} I(Y|X) &= I(f_{Y|X};\ f_Y) \\ &= E_X\ I\ (f_{Y|X};\ f_Y) \\ &= \int_{-\infty}^{\infty} I(f_{Y|X=x};\ f_Y)\ \ f_X(x)\, dx \end{aligned}$$

A fundamental fact is that

(**) $\qquad I(Y|X) = H(Y) - H(Y|X)$

Proof: $\qquad I\ (f_{Y|X=x};\ f_Y) = H(f_{Y|X=x};\ f_Y) - H(f_{Y|X=x})$

Take expectation with respect to X and verify that

$$\int_{-\infty}^{\infty} H(f_{Y|X=x};\ f_Y)\ f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{-\log f_Y(y)\} f_{X,Y}\ (x,y)\, dx\, dy = H(Y)$$

The most fundamental concept used in identifying models by estimating information is $I(Y|X_1;\ X_1,\ X_2)$, the information about Y in $X_2$ conditional on $X_1$; it is defined, by analogy with equation (**),

$$(***) \qquad I(Y|X_1;\, X_1,\, X_2) = H(f_{Y|X}) - H(f_{Y|X_1,\, X_2})$$

$$= H(Y|X_1) - H(Y|X_1,\, X_2).$$

A fundamental formula to evaluate $I(Y|X_1;\, X_1,\, X_2)$ is

$$(****) \qquad I(Y|X_1;\, X_1,\, X_2) = I(Y|X_1,\, X_2) - I(Y|X_1)$$

When X and Y are jointly normal random variables, $f_{Y|X=x}(y)$ is a normal distribution whose variance (which does not depend on x) is denoted $\Sigma(Y|X)$. The variance of Y is denoted $\Sigma(Y)$. The entropy and conditional entropy of Y are

$$H(Y) = \frac{1}{2} \log \Sigma(Y) + \frac{1}{2} (1 + \log 2\pi)$$

$$H(Y|X) = \frac{1}{2} \log \Sigma(Y|X) + \frac{1}{2} (1 + \log 2\pi)$$

The information about Y in X is written

$$I(Y|X) = -\frac{1}{2} \log \Sigma^{-1}(Y)\, \Sigma(Y|X)$$

When Y and X are jointly multivariate normal random vectors, let $\Sigma$ denote a covariance matrix. One can show that

$$I(Y|X) = (-\frac{1}{2}) \log \det \Sigma^{-1}(Y)\, \Sigma(Y|X)$$

$$= (-\frac{1}{2}) \text{ sum log eigenvalues } \Sigma^{-1}(Y)\, \Sigma(Y|X).$$

To make the foregoing formulas concrete, and to describe the general approach of this paper, consider the general problem of testing the hypothesis $H_o$: X and Y are independent. One could express $H_o$ in any one of the following equivalent ways:

$$H_o: \ f_{X,Y}(x,y) = f_X(x)\, f_Y(y) \text{ for all } x \text{ and } y;$$

$$H_o: f_{Y|X=x}(y) = f_Y(y) \text{ for all } x \text{ and } y;$$

$$H_o: I(f_{X,Y};\ f_X f_Y) = 0;$$

$$H_o: I(Y|X) = 0$$

The information approach to testing $H_o$ is to form an estimator $\hat{I}(Y|X)$ of $I(Y|X)$, and test whether it is significantly different from zero. One can distinguish several types of estimators of $I(Y|X)$: (a) fully parametric, (b) fully non-parametric; (c) functionally parametric which uses functional statistical inference smoothing techniques to estimate $I(Y|X)$ [see Woodfield (198

In this paper we consider only fully parametric estimators based on assuming multivariate normality of Y and X. When X and Y are bivariate normal with correlation coefficient $\rho$,

$$I(Y|X) = -\frac{1}{2} \log (1-\rho^2).$$

Given a random sample $(X_1, Y_1),\ldots,(X_n, Y_n)$ the maximum likelihood estimator of $I(Y|X)$ is

$$\hat{I}(Y|X) = -\frac{1}{2}\log(1-\hat{\rho}^2)$$

where $\hat{\rho}$ is the sample correlation coefficient. A test of $H_0$ based on $\hat{\rho}$ would reject $H_0$ at the 5% level of significance if $|\hat{\rho}|$ is greater than the threshold given in the following table:

| Sample Size n | Threshold for $\lvert\hat{\rho}\rvert$ | Threshold for $\hat{I}(Y\|X)$ |
|---|---|---|
| 20 | .444 | .11 |
| 40 | .312 | .05 |
| 50 | .279 | .04 |
| 80 | .220 | .025 |
| 100 | .197 | .02 |
| 150 | .160 | .013 |
| 200 | .139 | .01 |
| n | ? | 2/n |

In the foregoing table one sees a remarkable regularity in the 5% significance levels for the estimated information; they are approximately given by the simple formula 2/n. Test statistics based on entropy have 5% significance levels obeying the approximate rule m/n where n is the sample size and m is a constant which varies with the statistic used. At this time this perceived regularity is mainly an empirical fact; its theoretical basis is the conjecture that asymptotically $2n\,\hat{I}(Y|X)$ has a Chi-squared distribution with a suitable number m of degrees of freedom. If one transforms the 5% significance levels of the multiple correlation coefficient to significance levels for $I = -\frac{1}{2}\log(1-R^2)$, one discovers that the transformed critical values <u>approximately</u> obey the formula $(1 + k)/n$, where n is the sample size, and k is the number of regression variables. These empirical facts support the recommendation that statisticans should in their thinking replace $R^2$ by information I.

## 2. Information Formulation of Tests for White Noise

Let $\{Y(t),\ t=0, \pm 1,\ldots\}$ be a zero mean stationary Gaussian time series. The information about the value $Y(t)$ at time $t$ in the $m$ most recent values $Y(t-1),\ldots,Y(t-m)$ is denoted

$$I_m = I(Y(t)|\ Y(t-1),\ldots,Y(t-m))$$

It is more convenient to write henceforth

$$I_m = I(Y|Y_{-1},\ldots,Y_{-m}) \quad .$$

Introduce now the following notation for predictors (conditional expectations):

$$Y^{\mu,m}(t) = E[Y(t)|Y(t-1),\ldots,Y(t-m)] = (Y|Y_{-1},\ldots,Y_{-m})(t);$$

$$Y^{\nu,m}(t) = Y(t) - Y^{\mu,m}(t)$$

$$\sigma_m^2 = E[|Y^{\nu,m}(t)|^2] \div E[|Y(t)|^2]$$

$$= \Sigma(Y|Y_{-1},\ldots,Y_{-m})\ \Sigma^{-1}(Y)$$

The information $I_m$ about $Y$ in $Y_{-1},\ldots,Y_{-m}$ satisfies

$$I_m = -\frac{1}{2}\log \sigma_m^2$$

Next, let $Y^-$ denote the infinite past $Y(t-1), Y(t-2),\ldots$, and let

$$I_\infty = I(Y|Y^-).$$

One can show that

$$I_\infty = -\frac{1}{2}\log\sigma_\infty^2 = \left(-\frac{1}{2}\right)\int_0^1 \log f(\omega)\, d\omega$$

where $f(\omega)$ is the spectral density function of the time series $Y(t)$ satisfying

$$\rho(v) = E[Y(t)\,Y(t+v)] \div E[Y^2(t)]$$

$$= \int_0^1 \exp(2\pi i v\omega)\, f(\omega)\, d\omega, \quad v = 0, \pm 1, \ldots .$$

One of the powerful properties of information is that $I_\infty$ can be evaluated as a limit of $I_m$:

$$\lim_{m\to\infty} I_m = I_\infty .$$

The value of $I_\infty$ (in the Gaussian case, the value of $\sigma_\infty^2$) is used to classify the memory type of the time series as defined by Parzen (1981); a stationary (Gaussian) time series $Y(\cdot)$ is defined to be:

no memory if $I_\infty = 0$ $(\sigma_\infty^2 = 1)$ ;

short memory if $0 < I_\infty < \infty$ $(0 < \sigma_\infty^2 < 1)$;

long memory if $I_\infty = \infty$ $(\sigma_\infty^2 = 0)$ .

To estimate $I_m$, for m=1,2,..., and also $I_\infty$, from a sample Y(t), t=1,2,...,T, one uses the same estimators as if one were fitting an autoregressive scheme of order m to the time series:

$$Y(t) + \alpha_m(1)\, Y(t-1) + \ldots + \alpha_m(m)\, Y(t-m) = \varepsilon(t)$$

where $\varepsilon(t)$ is a white noise time series with variance denoted

$$\sigma_m^2 = E|\varepsilon(t)|^2 \doteq E|Y(t)|^2 .$$

We do not explicitly write the formulas for the estimators $\hat{\sigma}_m^2$ .

The hypothesis

$$H_o: Y(t) \text{ is white noise}$$

can be formulated in terms of information measures as

$$H_o: I_m = 0 \text{ for } m = 1,2,\ldots$$

For any fixed m to test the hypothesis that $I_m = 0$ one forms a test statistic of the form

$$\hat{I}_m = -\frac{1}{2} \log \hat{\sigma}_m^2$$

A 95% significance level for $\hat{I}_m$ seems to be approximately equivalent to one of the form

$$\hat{I}_m \leq \frac{m^*}{T}$$

where T is the time series sample size and m* is a suitable constant which depends on the order m (of the predictor) and the sample size T. Two widely used formulas for m* are [see Shibata (1981) for references]:

(1) m* = m, Akaike criterion;

(2) m* = m (log log T), Hannan-Quinn criterion.

The optimal value of m* for a given order m could be determined by Monte Carlo simulation. However we need a sequence of thresholds $\bar{I}_m$ so that the test region

$$\hat{I}_m \leq \bar{I}_m \quad \text{for } m = 1,2,...$$

provides an "optimum" test of the hypothesis that the time series is white noise. In choosing the critical values $\bar{I}_m$, one will undoubtedly use random walk theory since one can represent

$$\hat{I}_m = -\frac{1}{2} \log \hat{\sigma}_m^2 = \sum_{j=1}^{m} -\frac{1}{2} \log \{1-\hat{\rho}^2 (j|1,...,j-1)$$

where $\rho$ (j|1,...,j-1) is the partial correlation coefficient of Y(t) and Y(t-j) conditioned on Y(t-1),...,Y(t-(j-1)). The sample partial correlation coefficients $\hat{\rho}$(j|1,...,j-1) are asymptotically independent N(0,(1/n)) under the hypothesis $H_0$: Y(·) is white noise. The important work of Anderson (1971), p. 270, on the model order determination problem should be related to the random walk approach..

## 3. Information Formulation of ARMA Models

A white noise time series is characterized by the fact that the past has no information about the present. An autoregressive of order p, or AR(p), time series can be defined as one for which the most recent p values has as much information as the infinite past. In symbols, the following two hypotheses are equivalent:

$$H_o: Y(\cdot) \text{ is AR(p)},$$

$$H_o: \quad I(Y|Y_{-1},\ldots,Y_{-p};\ Y^{-}) = I_\infty - I_p = 0$$

An ARMA (p,q) scheme is usually defined by the representation

$$Y(t) + \alpha_p(1)\ Y(t-1) +\ldots+\alpha_p(p)\ Y(t-p)$$

$$= \varepsilon(t) + \beta_q(1)\ \varepsilon(t-1) +\ldots+\beta_q(q)\ \varepsilon(t-q)$$

where the polynomials

$$g_p(z) = 1 + \alpha_p(1)\ Z +\ldots+\alpha_p(p)\ z^p$$

$$h_q\ (z) = 1 + \beta_q(1)\ z +\ldots+ \beta_q(q)\ z^q$$

are chosen so that all their roots in the complex z-plane are in the region $\{z:|z|>1\}$ outside the unit circle.

To give an information characterization define the innovation time series

$$Y^{\nu}(t) = Y(t) - Y^{\mu}(t) = \lim_{m\to\infty} Y^{\nu,m}(t),$$

$$Y^{\mu}(t) = E[Y(t)|Y(t-1), Y(t-2),\ldots\ldots] = (Y|Y^{-})\ (t)$$

The following hypotheses can be shown to be equivalent:

$H_o$: $Y(\cdot)$ is ARMA $(p,q)$.

$$H_o\colon\ I(Y|Y_{-1},\ldots,Y_{-p},\ Y^{\nu}_{-1},\ldots,Y^{\nu}_{-q};\ Y^{-}) = 0$$

$$H_o\colon\ (Y|Y_{-1},\ldots,Y_{-p},\ Y^{\nu}_{-1},\ \ldots,\ Y^{\nu}_{-q})\ (t) = (Y|Y^{-})(t)$$

To compute the information one needs to compute the conditional variance $\Sigma(Y|Y_{-1},\ldots,Y_{-p},\ Y^{\nu}_{-1},\ldots,Y^{\nu}_{-q})$. To do this in practice we propose the following procedure:

1. Fit an AR($\hat{p}$) of order $\hat{p}$ determined by an order determination criterion.

2. Invert the AR($\hat{p}$) to form its MA($\infty$), infinite moving average representation,

$$Y(t) = Y^{\nu}(t) + \beta_1\ Y^{\nu}(t-1) + \beta_2\ Y^{\nu}(t-2) + \ldots$$

which is a non-parametric estimator of the MA($\infty$) representaiton. Note that

$$1 = \sigma^2_{\infty}\ \{1 + \beta_1^2 + \beta_2^2 + \ldots\}$$

and that the correlations $\rho(v) = \text{Corr}\ [Y(t), Y(t+v)]$ are estimated by

$$\rho(v) = \sigma_\infty^2 \{\beta_v + \beta_1\beta_{v+1} + \ldots\}$$

3. Form the joint covariance matrix of $Y(t)$, $Y(t-1)\ldots,Y(t-p)$, $Y^\nu(t-1),\ldots,Y^\nu(t-q)$ for suitable values of p and q. By using matrix sweep operators one can form the desired conditional variance $\sigma^2_{p,q} = \Sigma^{-1}(Y)$ $\Sigma(Y|Y_{-1},\ldots,Y_{-p}, Y^\nu_{-1}, \ldots,Y^\nu_{-q})$.

Note that

$$I(Y|Y_{-1},\ldots,Y_{-p}, Y^\nu_{-1},\ldots,Y^\nu_{-q}; Y^-) = I_\infty - I_{p,q},$$

$$I_{p,q} = -\frac{1}{2}\log \sigma^2_{p,q}$$

We illustrate this procedure by stating the conclusion for an ARMA(1,1):

$$I(Y|Y_{-1},Y^\nu_{-1};Y^-) = \frac{1}{2}\log\left\{\frac{1-\rho^2(1)}{\sigma^2_\infty} - \frac{\{\beta_1-\rho(1)\}^2}{1-\sigma^2_\infty}\right\}$$

One can verify that this information number equals 0 if the time series obeys any one of the schemes AR(1), MA(1), or ARMA(1,1). The information numbers for an AR(1) and MA(1) are respectively

$$I(Y|Y_{-1};Y^-) = \frac{1}{2}\log\left\{\frac{1-\rho^2(1)}{\sigma^2_\infty}\right\} \quad ;$$

$$I(Y|Y^\nu_{-1};Y^-) = \frac{1}{2}\log\left\{\frac{1}{\sigma^2_\infty} - \beta_1^2\right\}.$$

We do not discuss rigorously the method by which one chooses the best fitting ARMA (p,q). The method introduced by Akaike can be regarded as

computing for each p,q an estimator $\hat{I}_{p,q}$ of information from which one subtracts its significance level (a multiple of expected value) $\bar{I}_{p,q}$ under the hypothesis of white noise. Analogues of subset regression methods also seem to work in practice, and are used in our time series programs ARSPID and TIMESBOARD.

## 4. Multiple Time Series Model Identification

Let $Y = \{Y(t),\ t=0, \pm 1, \ldots\}$ be a multiple zero mean Gaussian stationary time series. One seeks to model $Y(t)$ in terms of its own past values, and values of multiple time series $X = \{X(t),\ t=0, \pm 1, \ldots\}$. A model begins with a representation

$$Y(t) = Y^{\mu}(t) + Y^{\nu}(t)$$

where $Y^{\mu}(t)$ is the linear predictor of $Y(t)$ given specified variables in the set $\{Y(t-1), Y(t-2), \ldots,; X(s),\ s = 0, \pm 1, \ldots\}$.

One always defines $Y^{\nu}(t) = Y(t) - Y^{\mu}(t)$. The probability law of the zero mean Gaussian multiple time series $\{Y^{\nu}(t),\ t=0, \pm 1, \ldots\}$ is described by the sequence of prediction error covariance matrices

$$\Sigma_{Y^{\nu}}(v) = E[Y^{\nu}(t)\{Y^{\nu}(t+v)\}^{*}]$$

where * denotes the complex conjugate of a matrix. The zero lag covariance $\Sigma_{Y^{\nu}}(0)$ is used in the evaluation of information. This matrix is written

$$\Sigma(Y|\text{predictor variables})$$

to indicate clearly which variables are used. We now describe various important information numbers and how they are computed (sample analogues of the following formulas are used for estimation). The information numbers we form are of the form $I(Y|X)$ or $I(Y|X_1;\ X_1,\ X_2)$ where $X$, $X_1$, $X_2$ are sets of predictor variables. $I(Y|X) = 0$ means that there is no significant

dependence of Y on the variables in X; $I(Y|X) > 0$ means that one can predict Y from the variables in X. $I(Y|X_1; X_1, X_2) = 0$ means that there is no information about Y in $X_2$ in addition to the information about Y in $X_1$. For each information number we list two hypotheses $H_0$ and $H_1$ which the information number can be used as a test statistic to distinguish. We write: $X^-$ to denote past X (the set X(t-1),...); $X^+$ to denote past and present X (the set X(t), X(t-1),...); X to denote all (past, present, and future) X (the set X(s), s = 0, $\pm$ 1, ...).

To decide which explanatory variables to use in modeling Y one computes estimators of the information numbers $I(Y|Y^-)$, $I(Y|X^-,Y^-)$, $I(Y|X^+,Y^-)$, $I(Y|X,Y^-)$, $I(Y|X)$ which one compares with their respective expected values to determine which information number most exceeds its expected or threshold values.

(1) $I(Y|Y^-)$, the information about Y in the infinite past of Y, is determined by computing (using Yule Walker equations) for p=1,2,...

$$I(Y|Y_{-1},\ldots,Y_{-p}) = (-\tfrac{1}{2}) \log \det \Sigma^{-1}(Y)\ \Sigma(Y|Y_{-1},\ldots,Y_{-p})$$

and determining an order $\hat{p}$ such that the value of the information about Y(t) in the p past values Y(t-1),...,Y(t-p) is used as an estimator of the information about Y(t) in Y(t-1), Y(t-2),... . This estimator statisfies the general formula

$$\log \det \Sigma(Y|Y^-) = \int_0^1 \log \det f_Y(\omega)\ d\omega$$

if the spectral density matrix of Y(·) is estimated by the autoregressive spectral density estimator of order $\hat{p}$.

For use in (5), we also compute at this stage $I(X|X^-)$.

(2) $I(Y|X^-,Y^-)$, the information about Y in the infinite past of X and Y, is determined by fitting multiple autoregressive schemes of order $p = 1,2,\ldots$ to the joint time series $\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix}$ which are used (for a suitable order p) to estimate the mean square prediction error matrices $\Sigma(X,Y|X^-,Y^-)$. It is represented as a partitioned matrix

$$\Sigma(X,Y|X^-,Y^-) = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

where $\Sigma_{XX} = \Sigma(X|X^-,Y^-)$, $\Sigma_{YY} = \Sigma(Y|X^-,Y^-)$, $\Sigma_{XY}$ is the conditional covariance matrix of X and Y, given $X^-$ and $Y^-$. Then

$$I(Y|X^-,Y^-) = (-\tfrac{1}{2}) \log \det \Sigma^{-1}(Y)\ \Sigma(Y|X^-,Y^-)$$

We also compute at this stage $I(X|X^-,Y^-)$ which is used in (5). The approximating autoregressive scheme is also used to estimate the spectral density matrix

$$f_{X,Y}(\omega) = \begin{bmatrix} f_{XX}(\omega) & f_{XY}(\omega) \\ f_{YX}(\omega) & f_{YY}(\omega) \end{bmatrix}$$

which is used in (3), and coherency $C(\omega) = f_{YY}^{-1}(\omega)\ f_{YX}(\omega)\ f_{XX}^{-1}(\omega) f_{XY}(\omega)$.

Several important identities can now be stated. The determinant of a partitioned matrix can be evaluated

$$\log\det \Sigma(X,Y|X^-,Y^-) = \log\det \Sigma_{XX} + \log\det \Sigma_{YY} - \Sigma_{YX}\,\Sigma_{XX}^{-1}\,\Sigma_{XY}$$

However $\Sigma_{XX} = \Sigma(X|X^-,Y^-)$, and Parzen (1969), p. 402 shows that

$$\Sigma(Y|X^+,Y^-) = \Sigma_{YY} - \Sigma_{YX}\,\Sigma_{XX}^{-1}\,\Sigma_{XY}$$

Thus we have the identity:

$$\text{(I)}\quad \log\det \Sigma(X,Y|X^-,Y^-) = \log\det \Sigma(X|X^-,Y^-) + \log\det \Sigma(Y|X^+,Y^-)$$

Next

$$\log\det\ f_{X,Y}(\omega) = \log\det f_{XX}(\omega) + \log\det f_{YY}(\omega) - f_{YX}(\omega)\,f_{XX}^{-1}(\omega)\,f_{XY}(\omega)$$

Integrating with respect to $\omega$ over $0\leq\omega\leq 1$, we obtain the identity

$$\text{(II)}\quad \log\det \Sigma(X,Y|X^-,Y^-) = \log\det \Sigma(X|X^-) + \log\det \Sigma(Y|X,Y^-)$$

since the spectral density of the error time series $(Y|X)^{\nu}(t) = Y(t)-(Y|X)(t)$ is

$$f_{(Y|X)^{\nu}}(\omega) = f_{YY}(\omega) - f_{YX}(\omega)\,f_{XX}^{-1}(\omega)\,f_{XY}(\omega)$$

Identities (I) and (II) play an important role below in stage (5); their importance may have been first pointed out by Geweke (1982), Theorem 1.

(3) $I(Y|X)$, the information about Y in all of X, is computed by

$$I(Y|X) = (-\tfrac{1}{2})\log\det \Sigma^{-1}(Y)\,\Sigma(Y|X)$$

where $\Sigma(Y|X) = \int_0^1 f_{YY}(\omega)\ \{I-C(\omega)\}d\omega$

$$= \int_0^1 \{f_{YY}(\omega) - f_{YX}(\omega)\ f_{XX}^{-1}(\omega)\ f_{XY}(\omega)\}\ d\omega$$

(4) $I(Y|X^+,Y^-)$, the information about Y in the past and present of X and the past of Y is given by

$$I(Y|X^+,Y^-) = (-\tfrac{1}{2}) \log \det \Sigma^{-1}(Y)\ \Sigma(Y|X^+,Y^-)$$

where $\Sigma(Y|X^+,Y^-) = \Sigma_{YY} - \Sigma_{YX}\ \Sigma_{XX}^{-1}\ \Sigma_{XY}$ in terms of the partitioned submatrices appearing in $\Sigma(X,Y|X^-,Y^-)$ computed in (2).

(5) $I(Y|X,Y^-)$, the information about Y in all of X and the past of Y, is computed in an ingenious manner developed by econometricians in their study of feedback measures [See Geweke (1982)]. First

$$I(Y|X,Y^-) = I(Y|Y^-) + I(Y|Y^-;X,Y^-)$$

Next

$$I(Y|Y^-;X,Y^-) = I(Y|Y^-;X^+,Y^-) + I(Y|X^+,Y^-;X,Y^-)$$

The first conditional information on the right hand side is computed

$$I(Y|Y^-;X^+,Y^-) = I(Y|X^+,Y^-) - I(Y|Y^-)$$

in terms of the information determined in (4) and (1) respectively. The second conditional information, defined by

$$I(Y|X^+,Y^-;X,Y^-) = I(Y|X,Y^-) - I(Y|X^+,Y^-) \quad ,$$

is computed by

$$I(Y|X^+,Y^-;X,Y^-) = I(X|X^-;X^-,Y^-)$$

(*****)

$$= I(X|X^-,Y^-) - I(X|X^-)$$

in terms of information computed in (2) and (1) respectively. A proof of equation (*****) is based on the identity

$$\log \det \Sigma(X,Y|X^-,Y^-)$$

$$= \log \det \Sigma(Y|X^+,Y^-) + \log \det \Sigma(X|X^-,Y^-)$$

$$= \log \det \Sigma(Y|X, Y^-) + \log \det \Sigma(X|X^-)$$

which follows from (I) and (II) in stage (2). Therefore

$$\log \det \Sigma(Y|X^+,Y^-) - \log \det \Sigma(Y|X,Y^-)$$

$$= \log \det \Sigma(X|X^-) - \log \det \Sigma(X|X^-,Y^-)$$

Summary. A method of summarizing the various information numbers is provided by reporting each of the terms in the following information decomposition:

$$I(Y|Y^-;X,Y^-) = I(Y|X,Y^-) - I(Y|Y^-) =$$

$$I(Y|Y^-;X^-,Y^-) + I(Y|X^-,Y^-;X^+,Y^-) + I(Y|X^+,Y^-;X,Y^-)$$

which enables one to construct the information numbers in (1), (2), (4), and (5). One also reports $I(Y|X)$ and $I(Y|X;X,Y^-)$.

The difference between measures of information is illuminated by expressing them when possible in spectral terms:

$$I(Y|Y^-;X,Y^-) = \int_0^1 (-\tfrac{1}{2}) \log \det \{I-C(\omega)\}\, d\omega,$$

$$I(Y|X;X,Y^-) = \tfrac{1}{2} \log \det \int_0^1 f_{YY}(\omega)\, \{I-C(\omega)\}\, d\omega$$

$$- \int_0^1 \tfrac{1}{2} \log \det f_{YY}(\omega)\, \{I-C(\omega)\}\, d\omega$$

Causality and Feedback. It should be noted that notions of feedback and causality studied by econometricians [see Gewerke (1982)] can be easily defined in terms of *information numbers*:

measure of linear dependence is $I(Y|Y^-;X,Y^-)$

measure of linear feedback from X to Y is $I(Y|Y^-;X^-,Y^-)$ ;

measure of instantaneous linear feedback is $I(Y|X^-,Y^-;X^+,Y^-)$.

5. Information Summary and Example

To summarize the relations between two multiple time series $X(\cdot)$ and $Y(\cdot)$ one estimates

I. Memory Measures

$I(X|X^-)$, $I(Y|Y^-)$

II. Feedback Measures

$I(X|X^-;X^-,Y^-)$, $I(Y|Y^-;X^-,Y^-)$, $I(Y|X^-,Y^-;X^+,Y^-)$

III. Information Increment Measures

$I(Y|Y^-;X^-,Y^-)$

$I(Y|Y^-;X^+,Y^-)$

$I(Y|Y^-;X,Y^-)$

$I(Y|X;X,Y^-)$

As an example, let us consider univariate time series X and Y which are given as Series J by Box and Jenkins (1970); X is gas furnace data, and Y is $CO_2$ in output gas. The time series sample size is T = 296. The means and standard deviations are given by

| | X | Y |
|---|---|---|
| Mean | -.057 | 53.51 |
| Standard deviation | 1.07 | 3.20 |

The ratio of standard deviations of Y to X is about 3; it can be regarded as a gain factor by which a change in X is multiplied into a change in Y.

The multiple covariances R(v) of the standardized time series (Y,X) are computed for v = 0,1,...,24; we list R(0), R(1), R(2), R(3), R(4), R(5):

$$\begin{bmatrix} 1.000 & -.485 \\ -.485 & 1.000 \end{bmatrix}, \begin{bmatrix} .971 & -.394 \\ -.598 & .953 \end{bmatrix}, \begin{bmatrix} .896 & -.329 \\ -.725 & .834 \end{bmatrix}$$

$$\begin{bmatrix} .793 & -.286 \\ -.843 & .682 \end{bmatrix}, \begin{bmatrix} .680 & -.260 \\ -.925 & .531 \end{bmatrix}, \begin{bmatrix} .575 & -.243 \\ -.950 & .408 \end{bmatrix}$$

The order determined AR schemes are: for X; order 6, $\Sigma(X|X^-) = .0302$; for Y: order 4, $\Sigma(Y|Y^-) = .0183$.

The order determined joint AR scheme for the standardized time series (Y,X) has order 4 and $\Sigma_{YY} = .0095$, $\Sigma_{XX} = .0306$, $\Sigma_{YX} = -.0021$. Then $\Sigma_{YY} - \Sigma_{YX}\,\Sigma_{XX}^{-1}\,\Sigma_{XY} = .0093$.

The spectral regression of standardized Y on all of standardized X has $\Sigma(Y|X) = .0618$.

The memory measures are (formulas apply to standardized X and Y)

$$I(X|X^-) = -.5 \log \Sigma(X|X^-) = 1.75,$$

$$I(Y|Y^-) = -.5 \log \Sigma(Y|Y^-) = 2.00;$$

one concludes that each time series has long memory.

*The feedback measures are*

$$I(Y|Y^-;X^-,Y^-) = .330$$

$$I(Y|X^-,Y^-;X^+,Y^-) = .008, \text{ not significantly different from zero,}$$

$$I(X|X^-;X^-,Y^-) = -.008, \text{ not significantly different from zero.}$$

The information increment measures are

$$I(Y|Y^-;X^-,Y^-) = .33$$

$$I(Y|Y^-;X^+,Y^-) = .33$$

$$I(Y|Y^-;X,Y^-) = .33$$

$$I(Y|X;X,Y^-) = .94$$

One interprets these measures to mean that adding $Y^-$ to X adds much more information than adding X to $Y^-$. Further adding $X^-$ to $Y^-$ is as informative as adding all X to $Y^-$.

## REFERENCES


Akaike, H. (1977). On entropy maximization principle, Applications of Statistics, P. R. Krishnaiah, ed., North-Holland: Amsterdam, 27-41.

Anderson, T. W. (1971). The Statistical Analysis of Time Series, Wiley: New York.

Box, G. E. P. and Jenkins, G. M. (1970). Time Series Analysis, Forecasting, and Control, San Francisco: Holden Day.

Geweke, J. (1982). The Measurement of Linear Dependence and Feedback Between Multiple Time Series, Journal of the American Statistical Association 77, 304-324.

IFAC (1982). Symposium on Identification and System Parameter Identification, Arlington, Virginia, June 7-11, 1982. Session on "New Ideas in System Identification" emphasizing information theory and entropy function approaches.

Parzen, E. (1967). Empirical multiple time series analysis, Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, edited by L. LeCam, and J. Neyman, University of California Press, I, 305-340.

Parzen, E. (1969). Multiple Time Series Modeling, Multivariate Analysis II, edited by P. Krishnaiah, Academic Press: New York, 389-409.

Parzen, E. (1981). Time Series Model Identification and Prediction Variance Horizon, Applied Time Series Analysis II, ed. D. Findley, Academic Press: New York, 415-447.

Parzen, E. (1982). Maximum Entropy Interpretation of Autoregressive Spectral Densities, Statistics and Probability Letters, 1, 2-6.

Parzen, E. and Newton, H. J. (1980). Multiple Time Series Modeling, II Multivariate Analysis - V, edited by P. Krishnaiah, North Holland: Amsterdam, 181-197.

Shibata, R. (1981). An optimal selection of regression variables. Biometrika 68, 45-54.

Stigler, S. M. (1982). A Modest Proposal: a New Standard for the Normal. The American Statistician 36, 137-138.

Woodfield, Terry J. (1982). Statistical Modeling of Bivariate Data. Ph.D. Thesis. Institute of Statistics, Texas A&M University.


END
FILMED